Probability Distributions

The distribution P_X of a random variable X is simply a probability measure which assigns probabilities to events on the real line. The distribution P_X answers questions of the form:

What is the probability that X lies in some subset F of the real line?

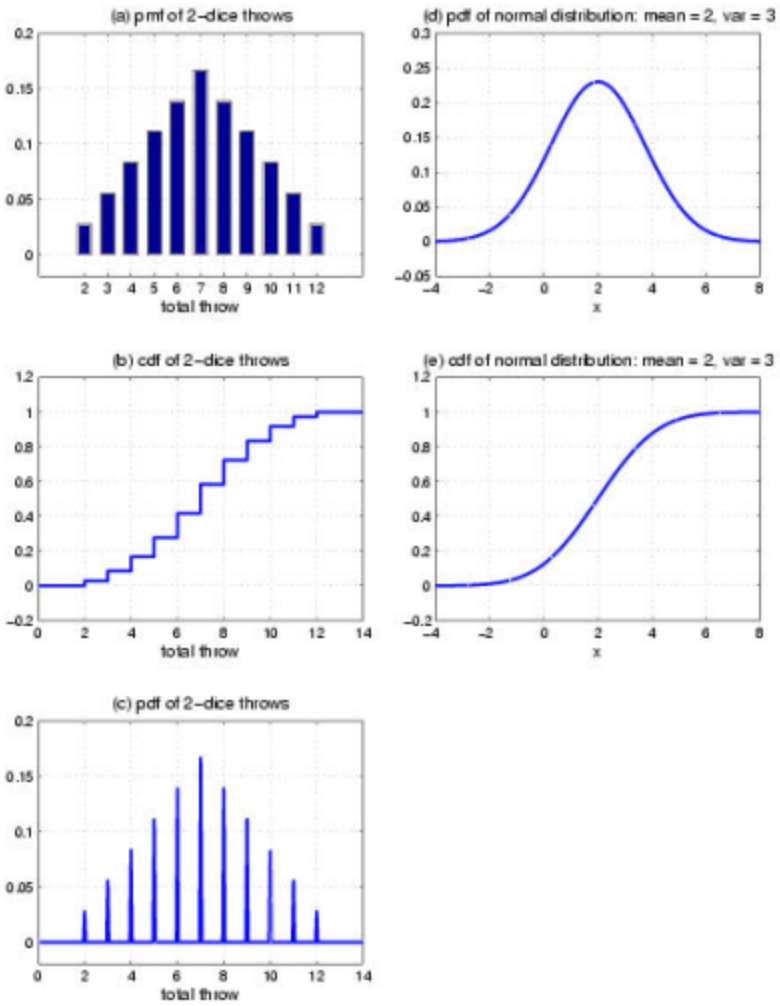
In practice we summarize P_X by its **Probability Mass Function - pmf** (for discrete variables only), **Probability Density Function - pdf** (mainly for continuous variables), or **Cumulative Distribution Function - cdf** (for either discrete or continuous variables).

Probability Mass Function (pmf)

Suppose the discrete random variable X can take a set of M real values $\{x_1, \ldots, x_M\}$, then the **pmf** is defined as:
Equation:

$$\begin{aligned} p_X(x_i) &= \Pr[X = x_i] \\ &= P_X(\{x_i\}) \end{aligned}$$

where $\sum_{i=1}^{M} p_X(x_i) = 1$. e.g. For a normal 6-sided die, $M = 6$ and $p_X(x_i) = \frac{1}{6}$. For a pair of dice being thrown, $M = 11$ and the pmf is as shown in (a) of [link].



Examples of pmfs, cdfs and pdfs: (a) to (c) for a discrete process, the sum of two dice; (d) and (e) for a continuous process with a normal or Gaussian distribution, whose mean = 2 and variance = 3.

Cumulative Distribution Function (cdf)

The **cdf** can describe discrete, continuous or mixed distributions of X and is defined as:

Equation:

$$\begin{aligned} F_X(x) &= \Pr[X \leq x] \\ &= P_X((-\infty, x]) \end{aligned}$$

For discrete X:
Equation:

$$F_X(x) = \sum_i \{p_X(x_i) | \, x_i \leq x\}$$

giving step-like cdfs as in the example of (b) of [link].

Properties follow directly from the Axioms of Probability:

1. $0 \leq F_X(x) \leq 1$
2. $F_X(-\infty) = 0$, $F_X(\infty) = 1$
3. $F_X(x)$ is non-decreasing as x increases
4. $\Pr[x_1 < X \leq x_2] = F_X(x_2) - F_X(x_1)$
5. $\Pr[X > x] = 1 - F_X(x)$

where there is no ambiguity we will often drop the subscript X and refer to the cdf as $F(x)$.

Probability Density Function (pdf)

The **pdf** of X is defined as the derivative of the cdf:
Equation:

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}\, x} F_X(x)$$

The pdf can also be interpreted in derivative form as $\delta(x) \to 0$:
Equation:

$$\begin{aligned} f_X(x)\delta(x) &= \Pr[x < X \leq x + \delta(x)] \\ &= F_X(x + \delta(x)) - F_X(x) \end{aligned}$$

For a discrete random variable with pmf given by $p_X(x_i)$:
Equation:

$$f_X(x) = \sum_{i=1}^{M} p_X(x_i)\delta(x - x_i)$$

An example of the pdf of the 2-dice discrete random process is shown in (c) of [link]. (Strictly the delta functions should extend vertically to infinity, but we show them only reaching the values of their areas, $p_X(x_i)$.)

The pdf and cdf of a continuous distribution (in this case the **normal** or **Gaussian** distribution) are shown in (d) and (e) of [link].

Note: The cdf is the integral of the pdf and should always go from zero to unity for a valid probability distribution.

Properties of pdfs:

1. $f_X(x) \geq 0$
2. $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}\,x = 1$
3. $F_X(x) = \int_{-\infty}^{x} f_X(\alpha)\,\mathrm{d}\,\alpha$
4. $\Pr[x_1 < X \leq x_2] = \int_{x_1}^{x_2} f_X(\alpha)\,\mathrm{d}\,\alpha$

As for the cdf, we will often drop the subscript X and refer simply to $f(x)$ when no confusion can arise.

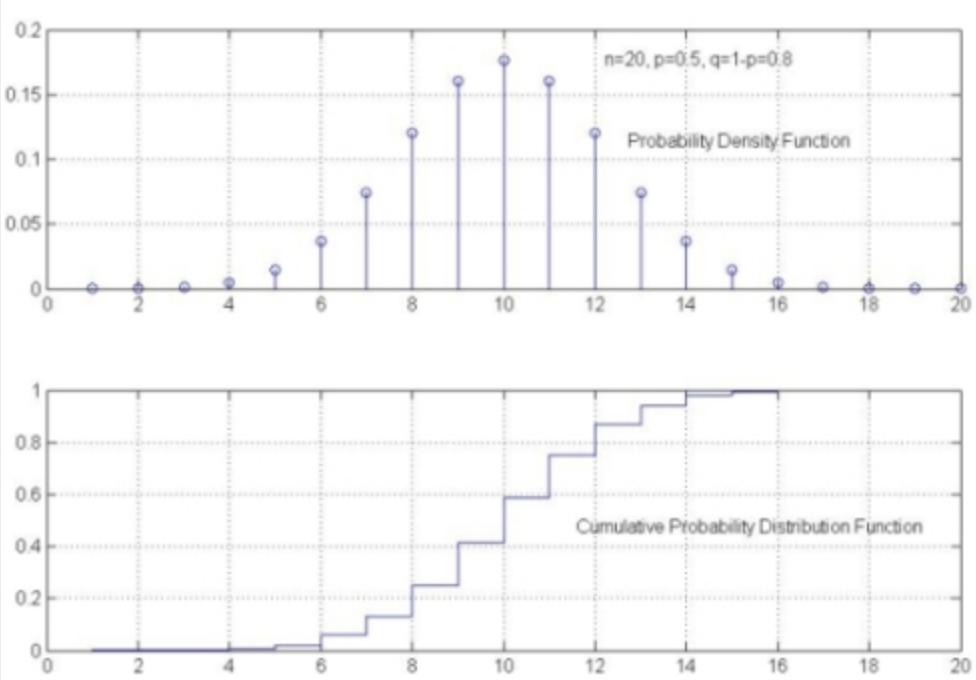
What is a probability distribution function?

A mathematical function can be used to model the frequencies and probabilities of occurrences over time. A discrete probability distribution function associates a list of probabilities with each possible value of a discrete random variable. The probability distribution function is thus used to model the probabilities of a discrete random variable and is also known as a **probability mass function**. The probabilities of a continuous random variable are modelled using continuous distribution functions, also known as **probability density functions** (pdf's).

The following are particularly important forms of the probability distribution function.

- Binomial distribution function. Used to model experiments with only two possible outcomes.
- Poisson distribution function. Used to model experiments with high degrees of uncertainty.
- Normal distribution. Used to model continuous distributions.

Example:
This discrete probability density function models experiments that have only two possible outcomes. The probability of success is **p** and the probability of failure is **q**=1-**p**. The pdf models the probability that we will observe **r** sucesses and **n**-**r** failures in a total of **n**-trials.



Graph of the probability distribution function and the cumulative probability distribution function (redrawn from http://www.engr.udayton.edu/faculty/mdaniels/htm315/Functions.htm using matlab)

Exercise:

Problem:

From the example above, what is the probability that in 20-trials there are exactly six successes?

Solution:

The probability that there are exactly six successes is **0.04**

References:

1. Random Variables and their Probability Density and Distribution Functions, http://www.engr.udayton.edu/faculty/mdaniels/htm315/Functions.htm (last accessed February 2006)
2. NCAR Advanced Study Program http://www.asp.ucar.edu (last accessed February 2006)

Co-Author: Mookho Tsilo

Signal Parameter Estimation

One extension of parametric estimation theory necessary for its application to array processing is the estimation of signal parameters. We assume that we observe a signal $s(l, \theta)$, whose characteristics are known save a few parameters θ, in the presence of noise. Signal parameters, such as amplitude, time origin, and frequency if the signal is sinusoidal, must be determined in some way. In many cases of interest, we would find it difficult to justify a particular form for the unknown parameters' a priori density. Because of such uncertainties, the minimum mean-squared error and maximum a posteriori estimators **cannot** be used in many cases. The minimum mean-squared error **linear** estimator does not require this density, but it is most fruitfully used when the unknown parameter appears in the problem in a linear fashion (such as signal amplitude as we shall see).

Linear Minimum Mean-Squared Error Estimator

The only parameter that is linearly related to a signal is the amplitude. Consider, therefore, the problem where the observations at an array's output are modeled as
Equation:

$$\forall l, l \in \{0, \ldots, L-1\} : (r(l) = \theta s(l) + n(l))$$

The signal waveform $s(l)$ is known and its energy normalized to be unity ($\sum_l s^2(l) = 1$). The linear estimate of the signal's amplitude is assumed to be of the form $\widehat{\boldsymbol{\theta}} = \sum_l h(l) r(l)$, where $h(l)$ minimizes the mean-squared error. To use the Orthogonality Principle expressed by this equation, an inner product must be defined for scalars. Little choice avails itself but multiplication as the inner product of two scalars. The Orthogonality Principle states that the estimation error must be orthogonal to all linear transformations defining the kind of estimator being sought.

$$\forall h(\cdot) : \left(E\left[\left(\sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l) r(l) - \theta \right) \sum_{k=0}^{L-1} h(k) r(k) \right] = 0 \right)$$

Manipulating this equation to make the universality constraint more transparent results in

$$\forall h(\cdot) : \left(\sum_{k=0}^{L-1} h(k) E\left[\left(\sum_{l=0}^{L-1} h_{\text{LIN}}(l) r(l) - \theta \right) r(k) \right] = 0 \right)$$

Written in this way, the expected value must be 0 for each value of k to satisfy the constraint. Thus, the quantity $h_{\text{LIN}}(\cdot)$ of the estimator of the signal's amplitude must satisfy

$$\forall k : \left(\sum_{l=0}^{L-1} h_{\text{LIN}}(l) E[r(l) r(k)] = E[\theta r(k)] \right)$$

Assuming that the signal's amplitude has zero mean and is statistically independent of the zero-mean noise, the expected values in this equation are given by

$$E[r(l) r(k)] = {\sigma_\theta}^2 s(l) s(k) + K_n(k, l)$$

$$E[\theta r(k)] = {\sigma_\theta}^2 s(k)$$

where $K_n(k, l)$ is the covariance function of the noise. The equation that must be solved for the unit-sample response $h_{\text{LIN}}(\cdot)$ of the optimal linear MMSE estimator of signal amplitude becomes

Equation:

$$\forall k : \left(\sum_{l=0}^{L-1} h_{\text{LIN}}(l) K_n(k, l) = {\sigma_\theta}^2 s(k) \left(1 - \sum_{l=0}^{L(1)} h_{\text{LIN}}(l) s(l) \right) \right)$$

This equation is easily solved once phrased in matrix notation. Letting K_n denote the covariance matrix of the noise, $\boldsymbol{s}$ the signal vector, and h_{LIN} the vector of coefficients, this equation becomes

$$K_n h_{\text{LIN}} = {\sigma_\theta}^2 \left(1 - \boldsymbol{s}^T h_{\text{LIN}} \right) \boldsymbol{s}$$

The matched filter for colored-noise problems consisted of the dot product between the vector of observations and ${K_n}^{-1}\boldsymbol{s}$ (see the detector result). Assume that the solution to the linear estimation problem is proportional to the detection theoretical one: $h_{\text{LIN}} = c{K_n}^{-1}\boldsymbol{s}$, where c is a scalar constant. This proposed solution satisfies the equation; the MMSE estimate of signal amplitude corresponds to applying a matched filter to the observations with **Equation:**

$$h_{\text{LIN}} = \frac{{\sigma_\theta}^2}{1 + {\sigma_\theta}^2 \boldsymbol{s}^T {K_n}^{-1} \boldsymbol{s}} {K_n}^{-1} \boldsymbol{s}$$

The mean-squared estimation error of signal amplitude is given by

$$E\left[\varepsilon^2\right] = {\sigma_\theta}^2 - E\left[\theta \sum_{l=0}^{L-1} h_{\text{LIN}}(l) r(l)\right]$$

Substituting the vector expression for h_{LIN} yields the result that the mean-squared estimation error equals the proportionality constant c defined earlier.

$$E\left[\varepsilon^2\right] = \frac{{\sigma_\theta}^2}{1 + {\sigma_\theta}^2 \boldsymbol{s}^T {K_n}^{-1} \boldsymbol{s}}$$

Thus, the linear filter that produces the optimal estimate of signal amplitude is equivalent to the matched filter used to detect the signal's presence. We have found this situation to occur when estimates of unknown parameters are needed to solve the detection problem (see Detection in the Presence of Uncertainties). If we had not assumed the noise to be Gaussian, however, this detection-theoretic result would be different, but the estimator would be unchanged. To repeat, this invariance occurs because the linear MMSE estimator requires **no** assumptions on the noise's amplitude characteristics.

Example:

Let the noise be white so that its covariance matrix is proportional to the identity matrix ($K_n = {\sigma_n}^2 I$). The weighting factor in the minimum mean-squared error linear estimator is proportional to the signal waveform.

$$h_{\text{LIN}}(l) = \frac{{\sigma_\theta}^2}{{\sigma_n}^2 + {\sigma_\theta}^2} s(l)$$

$$\hat{\theta}_{\text{LIN}} = \frac{{\sigma_\theta}^2}{{\sigma_n}^2 + {\sigma_\theta}^2} \sum_{l=0}^{L-1} s(l) r(l)$$

This proportionality constant depends only on the relative variances of the noise and the parameter. **If** the noise variance can be considered to be much smaller than the a priori variance of the amplitude, then this constant does not depend on these variances and equals unity. Otherwise, the variances must be known.
We find the mean-squared estimation error to be

$$E\left[\varepsilon^2\right] = \frac{{\sigma_\theta}^2}{1 + \frac{{\sigma_\theta}^2}{{\sigma_n}^2}}$$

This error is significantly reduced from its nominal value ${\sigma_\theta}^2$ only when the variance of the noise is small compared with the a priori variance of the amplitude. Otherwise, this admittedly optimum amplitude estimate performs poorly, and we might as well as have ignored the data and "guessed" that the amplitude was zero[footnote].
In other words, the problem is difficult in this case.

Linear Estimators

We derived the minimum mean-squared error estimator in the previous section with no constraint on the form of the estimator. Depending on the problem, the computations could be a linear function of the observations (which is always the case in Gaussian problems) or nonlinear. Deriving this estimator is often difficult, which limits its application. We consider here a variation of MMSE estimation by constraining the estimator to be linear while minimizing the mean-squared estimation error. Such **linear estimators** may not be optimum; the conditional expected value may be nonlinear and it **always** has the smallest mean-squared error. Despite this occasional performance deficit, linear estimators have well-understood properties, they interact will with other signal processing algorithms because of linearity, and they can always be derived, no matter what the problem.

Let the parameter estimate $\hat{\theta}(\boldsymbol{r})$ be expressed as $\mathscr{L}(\boldsymbol{r})$ where $\mathscr{L}(\cdot)$ is a linear operator: $\mathscr{L}(a_1 r_1 + a_2 r_2) = a_1 \mathscr{L}(r_1) + a_2 \mathscr{L}(r_2)$ where a_1, a_2 are scalars. Although all estimators of this form are obviously linear, the term **linear estimator** denotes that member of this family that minimizes the mean-squared error.
Equation:

$$\underset{\mathscr{L}(\boldsymbol{r})}{\operatorname{argmin}} \; E\left[\varepsilon^T \varepsilon\right] = \hat{\theta}_{\text{LIN}}(\boldsymbol{r})$$

Because of the transformation's linearity, the theory of linear vector spaces can be fruitfully used to derive the estimator and to specify its properties. One result of that theoretical framework is the well-known **Orthogonality Principle** (Papoulis, pp. 407-414) The linear estimator is that particular linear transformation that yields an estimation error orthogonal to all linear transformations of the data. The orthogonality of the error to **all** linear transformations is termed the **universality constraint.** This principle provides us not only with a formal definition of the linear estimator but also with the mechanism to derive it. To demonstrate this intriguing result, let $\langle \cdot, \cdot \rangle$ denote the absract inner product between two vectors and $\| \cdot \|$ the associated norm.
Equation:

$$(\| \boldsymbol{x} \|)^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$$

For example, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are each column matrices having only one column,[footnote] their inner product might be defined as $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \boldsymbol{x}^T \boldsymbol{y}$. Thus, the linear estimator as defined by the Orthogonality Principle must satisfy
Equation:

$$\forall \text{ for all linear transformations } \mathscr{L}(\cdot) : \left(E\left[\left\langle \hat{\theta}_{\text{LIN}}(\boldsymbol{r}) - \theta, \mathscr{L}(\boldsymbol{r}) \right\rangle \right] = 0 \right)$$

To see that this principle produces the MMSE linear estimator, we express the mean-squared estimation error $E\left[\varepsilon^T \varepsilon\right] = E\left[(\| \varepsilon \|)^2 \right]$ for **any** choice of linear estimator $\hat{\theta}$ as
Equation:

$$
\begin{aligned}
E\left[\left(\parallel \hat{\theta} - \theta \parallel\right)^2\right] &= E\left[\left(\parallel \hat{\theta}_{\text{LIN}} - \theta - \left(\hat{\theta}_{\text{LIN}} - \hat{\theta}\right) \parallel\right)^2\right] \\
&= E\left[\left(\parallel \hat{\theta}_{\text{LIN}} - \theta \parallel\right)^2\right] + E\left[\left(\parallel \hat{\theta}_{\text{LIN}} - \hat{\theta} \parallel\right)^2\right] - 2E\left[\left\langle \hat{\theta}_{\text{LIN}} - \theta, \hat{\theta}_{\text{LIN}} - \hat{\theta} \right\rangle\right]
\end{aligned}
$$

As $\hat{\theta}_{\text{LIN}} - \hat{\theta}$ is the difference of two linear transformations, it too is linear and is orthogonal to the estimation error resulting from $\hat{\theta}_{\text{LIN}}$. As a result, the last term is zero and the mean-squared estimation error is the sum of two squared norms, each of which is, of course, nonnegative. Only the second norm varies with estimator choice; we minimize the mean-squared estimation error by choosing the estimator $\hat{\theta}$ to be the estimator $\hat{\theta}_{\text{LIN}}$, which sets the second term to zero.
There is a confusion as to what a vector it. "Matricies having one column" are colloquially termed vectors as are the field quantities such as electric and magnetic fields. "Vectors" and their associated inner products are taken to be much more general mathematical objects than these. Hence the prose in this section is rather contorted.

The estimation error for the minimum mean-squared linear estimator can be calculated to some degree without knowledge of the form of the estimator. The mean-squared estimation error is given by
Equation:

$$
\begin{aligned}
E\left[\left(\parallel \hat{\theta}_{\text{LIN}} - \theta \parallel\right)^2\right] &= E\left[\left\langle \hat{\theta}_{\text{LIN}} - \theta, \hat{\theta}_{\text{LIN}} - \theta \right\rangle\right] \\
&= E\left[\left\langle \hat{\theta}_{\text{LIN}} - \theta, \hat{\theta}_{\text{LIN}} \right\rangle\right] + E\left[\left\langle \hat{\theta}_{\text{LIN}} - \theta, -\theta \right\rangle\right]
\end{aligned}
$$

The first term is zero because of the Orthogonality Principle. Rewriting the second term yields a general expression for the MMSE linear estimator's mean-squared error.
Equation:

$$
E\left[\left(\parallel \varepsilon \parallel\right)^2\right] = E\left[\left(\parallel \theta \parallel\right)^2\right] - E\left[\left\langle \hat{\theta}_{\text{LIN}}, \theta \right\rangle\right]
$$

This error is the difference of two terms. The first, the mean-squared value of the parameter, represents the largest value that the estimation error can be for any reasonable estimator. That error can be obtained by the estimator that ignores the data and has a value of zero. The second term reduces this maximum error and represents the degree to which the estimate and the parameter agree on the average.

Note that the definition of the minimum mean-squared error **linear** estimator makes no explicit assumptions about the parameter estimation problem being solved. This property makes this kind of estimator attractive in many applications where neither the a priori density of the parameter vector nor the density of the observations is known precisely. Linear transformations, however, are homogeneous: A zero-values input yields a zero output. Thus, the linear estimator is especially pertinent to those problems where the expected value of the parameter is zero. If the expected value is nonzero, the linear estimator would not necessarily yield the best result (See this problem)

Example:
Express the first example in vector notation so that the observation vector is written as

$$\boldsymbol{r} = A\theta + \boldsymbol{n}$$

where the matrix A has the form $A = (1 \ldots 1)^T$. The expected value of the parameter is zero. The linear estimator has the form $\hat{\theta}_{\text{LIN}} = L\boldsymbol{r}$, where L is a $1 \times L$ matrix. The orthogonality Principle states that the linear estimator satisfies

$$\forall \text{for all } 1 \times L \text{ matricies } M : \left(E\left[(L\boldsymbol{r} - \theta)^T M \boldsymbol{r} \right] = 0 \right)$$

To use the Orthogonality Principle to derive an equation implicitly specifying the linear estimator, the "for all linear transformations" phrase must be interpreted. Usually the quantity specifying the linear transformation must be removed from the constraining inner product by imposing a very stringent but equivalent condition. In this example, this phrase becomes one about matrices. The elements of the matrix M can be such that each element of the observation vector multiplies each element of the estimation error. Thus, in this problem the Othogonality Principle means that the expected value of the matrix consisting of all pairwise priducts of these elements must be zero.

$$E\left[(L\boldsymbol{r} - \theta)\boldsymbol{r}^T \right] = 0$$

Thus, two terms must equal each other: $E\left[L\boldsymbol{r}\boldsymbol{r}^T \right] = E\left[\theta \boldsymbol{r}^T \right]$. The second term equals $E\left[\theta^2 \right] A^T$ as the additive noise and the parameter are assumed to be statistically independent quantities. The quantity $E\left[\boldsymbol{r}\boldsymbol{r}^T \right]$ in the first term is the correlation matrix of the observations, which is given by $AA^T E\left[\theta^2 \right] + K_n$. Here, K_n is the noise covariance matrix, and $E\left[\theta^2 \right]$ is the parameter's variance. The quantity AA^T is a $L \times L$ matrix with each element equaling 1. The noise vector has independent components; the covariance matrix thus equals ${\sigma_n}^2 I$. The equation that L must satisfy is therefore given by

$$(L_1 \quad \cdots \quad L_L) \begin{pmatrix} {\sigma_n}^2 + {\sigma_\theta}^2 & {\sigma_\theta}^2 & \cdots & {\sigma_\theta}^2 \\ {\sigma_\theta}^2 & {\sigma_n}^2 + {\sigma_\theta}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & {\sigma_\theta}^2 \\ {\sigma_\theta}^2 & \cdots & {\sigma_\theta}^2 & {\sigma_n}^2 + {\sigma_\theta}^2 \end{pmatrix} = ({\sigma_\theta}^2 \quad \cdots \quad {\sigma_\theta}^2)$$

The components of L are equal and are given by $L_i = \frac{{\sigma_\theta}^2}{{\sigma_n}^2 + L{\sigma_\theta}^2}$. Thus, the minimum mean-squared error linear estimator has the form

$$\hat{\theta}_{\text{LIN}}(\boldsymbol{r}) = \frac{{\sigma_\theta}^2}{{\sigma_\theta}^2 + \frac{{\sigma_n}^2}{L}} \frac{1}{L} \sum_{ll} r(l)$$

Note that this result equals the minimum mean-squared error estimate derived earlier under the condition that $E[\theta] = 0$. Mean-squared error, linear estimators, and Gaussian problems are intimately related to each other. The linear minimum mean-squared error solution to a problem is optimal if the underlying distributions are Gaussian.